DOCUMENT RESUME

ED 266 145 TM 850 462

AUTHOR Angoff, William H.
TITLE Between-Groups Experimental Design for Studies of Rights and Formula Scoring.
PUB DATE 3 Apr 85
NOTE 9p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 31- April 3, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; *Guessing (Tests); Research Design; *Scoring Formulas; Secondary Education; Testing; *Test Interpretation; Test Items; *Test Theory
IDENTIFIERS *Rights and Formula Scoring

ABSTRACT

This paper points out that there are certain generalizations about directions for guessing and methods of scoring that require that data be derived from random groups design. It supports the viewpoint that it is neither sufficient nor appropriate to make such generalizations on the basis of an analysis of scores obtained from the answer sheets of only one group of students. An earlier study, using the principle of separate groups for separate instructions, was conducted by administering the SAT-Verbal to four randomly equivalent groups of students. It was found that if two groups were given the test with different directions, those given the test with Rights directions guessed more often and earned significantly higher Rights scores. When their papers were rescored by Formula, the Formula scores came much closer together. The two types of administration-and-scoring were about equally reliable. The correlation between Rights scores on one test and Formula scores on another was about the same as the correlation between Rights and Rights or between Formula and Formula. The approach that an examinee takes in responding to test questions is very much affected by the instructions he or she receives. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED266145

1/15/86 Revision

Between-Groups Experimental Design for Studies of Rights and Formula Scoring

William H. Angoff
Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. Weidenmiller
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

NCME Meetings
Chicago, Illinois
April 3, 1985

TM 850 462

The essential point of this paper is that there are certain generalizations about directions for guessing and methods of scoring that require that data be derived from random groups designs--that is to say, from the assignment of different random groups to different experimental treatments. It is neither sufficient nor appropriate to make such generalizations on the basis of an analysis of the scores obtained from the answer sheets of only one group of students. What has caused some confusions in the design of such studies, I suspect, is that the term, "Formula-Scoring" and the term, "Rights-Scoring," have been used variously, sometimes to refer to directions for administering the test, and sometimes to refer to the scoring procedure itself, and these are obviously different.

We have heard some researchers conclude, for example, that the difference in effect between Rights and Formula scores is trivial because they correlate so highly. Indeed, we hear it argued that if there are no omitted items, the correlation between them is perfect. A similar logic has led to the conclusion that since the usual Formula score, in which the Rights are diminished by a fraction of the number of Wrongs, is perfectly correlated with the Formula in which the Rights are augmented by a fraction of the Omits, there is no effective difference between the two. The defect in this conclusion derives from the fact that the two scores are based on the same responses to the same items, administered under one set of conditions, and leading, for the most part, to the adoption by the students of one particular guessing strategy. As a consequence, the scores lack the experimental independence that would normally permit one to draw such conclusions. The observed correlations are, in effect, spurious correlations--i.e., to a large extent, self-correlations. Clearly,

the conclusions are necessarily restricted only to the behavior of two methods of scoring; they say nothing about differences in the behavior of examinees when examined under two conditions of administration.

A similar difficulty in interpreting the results of analyses appears in the experiment in which students are asked to work their way through the items of a test with the knowledge that they will be penalized for wrong answers, then asked to change to differently colored pencils and instructed to guess at the answers to the items that were left unmarked on the first run-through. The two scores, one based on Formula directions, the other based on Rights directions, are then correlated and the conclusion is drawn that since they rank order the students in essentially the same way, there is no effective difference between them. Here too, of course, the scores are based to a considerable extent on the same responses, and the correlation between them is spurious.

Sometimes we find studies in the literature that appear to ignore the spuriousness. There is a frequently cited article, for example, (Sherriffs and Boomer, 1954), in which the authors correlated both the Rights scores and the Formula scores on a test administered with Formula directions against an outside criterion--high schoool GPA. The authors reported that the two scores correlated very highly with each other (.96) and about equally well with the criterion (.48 and .49), and left it to the reader to conclude that although the mean Rights score was higher than the mean Formula score, there was little difference between them. In fact, there was little difference between them. The two variables were actually very nearly the same variable. The upshot is that this part of their study contributed very little to our understanding of the problem.

Put this another way: It is not to say that the statement these researchers made was necessarily incorrect. It is only that the evidence for the statement--the inference they drew--was not warranted. An appropriate design which would enable one to make, or not to make, this inference is to separate the entire group into random halves, give one half Rights directions, the other half Formula directions, score each group in accordance with the directions they received, and correlate the scores for each group with the criterion. An examination of those correlations would lead one appropriately to consider the validity of the inference.

There are probably several studies in the literature--or at least, I hope there are--in which this type of design was used. One of these is by Traub, Hambleton, and Singh (1969), who administered two parallel forms of a test to four randomly chosen groups of examinees under four separate types of guessing instructions. This design permitted them to make comparisons within and across the groups, but with appropriate care to avoid the confusion between real and artificial, or spurious, similarities. In another study, carried out about four years ago, one which made use of the same principle of separate groups for separate instructions, William Schrader and I (Angoff and Schrader, 1984) administered the SAT-Verbal to four randomly equivalent groups of students in a plan described in your handout. With this arrangement, we were afforded an excellent opportunity to study some of the characteristics of Rights and Formula directions and Rights and Formula scores. For example, we found the following:

1. As expected, if two equivalent groups are given the test with different directions, those who are given the test with Rights directions

guess more often (as they are instructed to do), and earn significantly higher Rights scores than do those who are given the test with Formula directions. This, of course, comes as no surprise. What is surprising is that when their papers were both rescored by Formula, the Formula scores came much closer together--and, as we also found in another phase of the study, were virtually identical. Traub, Hambleton, and Singh, by the way, found the same result, but their paper did not highlight it. This finding suggests contrary to some claims, that Formula scoring does not give an advantage to those who are willing to take a risk and guess, nor does it confirm the claim that, overall, examinees rely on valid partial information when they guess. Another way of expressing this claim that some investigators make, and for which we find contrary data, is that people actually know more than they think they do, and that when they guess at an item, they do better than chance. What our data tell us is that this may possibly be true on some items, and it may possibly be true for some people--we have no data on this--but over all items and all people it is very nearly as often not the case. In fact, it is about four times as likely to be wrong as it is to be right.

2. As you know, some authors assert that Formula-scoring is more reliable; others claim that Rights-scoring is more reliable. In our study, the two types of administration-and-scoring (i.e., where the scoring was consistent with the directions for administration) were about equally reliable. Rights scores based on a Rights-administered test are about as reliable as Formula scores based on a Formula-administered test. This result was observed not only when the reliability was calculated by Kuder-Richardson (20) but also when calculated by parallel-forms procedure.

3. The correlation between Rights scores on one test and Formula scores on another is about the same--only very slightly lower--than the correlation between Rights and Rights or between Formula and Formula. This finding supports the conclusion that Rights and Formula scores yield very nearly the same information, but here, I would submit, the evidence for the finding is defensible. At least, it is based on experimentally independent observations.

4. The approach that an examinee takes in responding to test questions is very much affected by the instructions he or she receives. We have always known this to be true. What the study also shows is that examinees apparently find no difficulty in shifting and changing their strategies abruptly as required, and that they do so apparently with ease.

There were other detailed findings of interest in this study, but the point here is not so much the findings per se but the fact that the inferences based on these findings could not have been made without a random groups design. Rescoring a group of answer sheets for examinees who have taken a test under conditions inconsistent with that scoring method must be taken literally for what it is. Obviously, no rescoring can alter the fact that the examinees had taken the test and conducted themselves while taking it in a manner appropriate to a particular set of directions, and that at the time of the rescoring that behavior cannot be changed. Rights scores and Formula scores based on the same administration, given under one set of guessing directions are not independent observations, and should not be treated as such.

REFERENCES

Angoff, W. H. and Schrader, W. B. (1984). A study of hypotheses basic to the use of rights and formula scores. Journal of Educational Measurement, 21, 1-17.

Sherriffs, A. C. and Boomer, D. C. (1954). Who is penalized by guessing? Journal of Educational Psychology, 45, 81-90.

Traub, R. E., Hambleton, R. K., and Singh, B. (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. Educational and Psychological Measurement, 29, 847-861.

BEST COPY AVAILABLE

Guessing Instructions

for the Administration of the SAT-Verbal

	Section I	Section II
Group 1	Rights	Rights
Group 2	Rights	Formula
Group 3	Formula	Rights
Group 4	Formula	Formula

Note: The SAT-Verbal is a one-hour test, administered in two separately-timed half-hour sections designed to be parallel in content and difficulty, although slightly different in length: Section I contain 45 5-choice items; Section II contains 40 5-choice items. Both sections are operationally administered with formula directions and scored by the formula, R-W/4.